Understanding Dataset Difficulty with $\mathcal{V}$-Usable Information

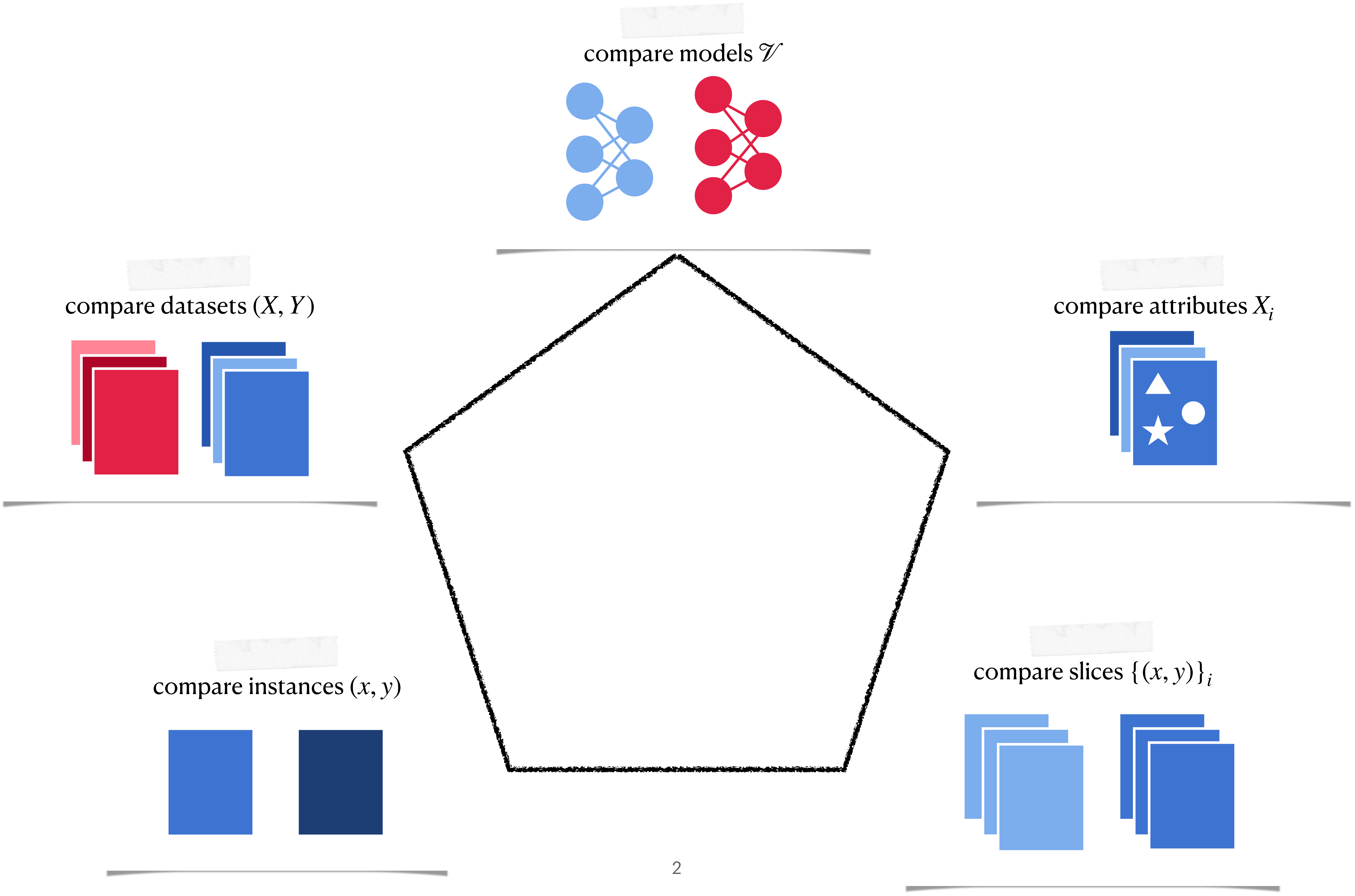
ICML 2022

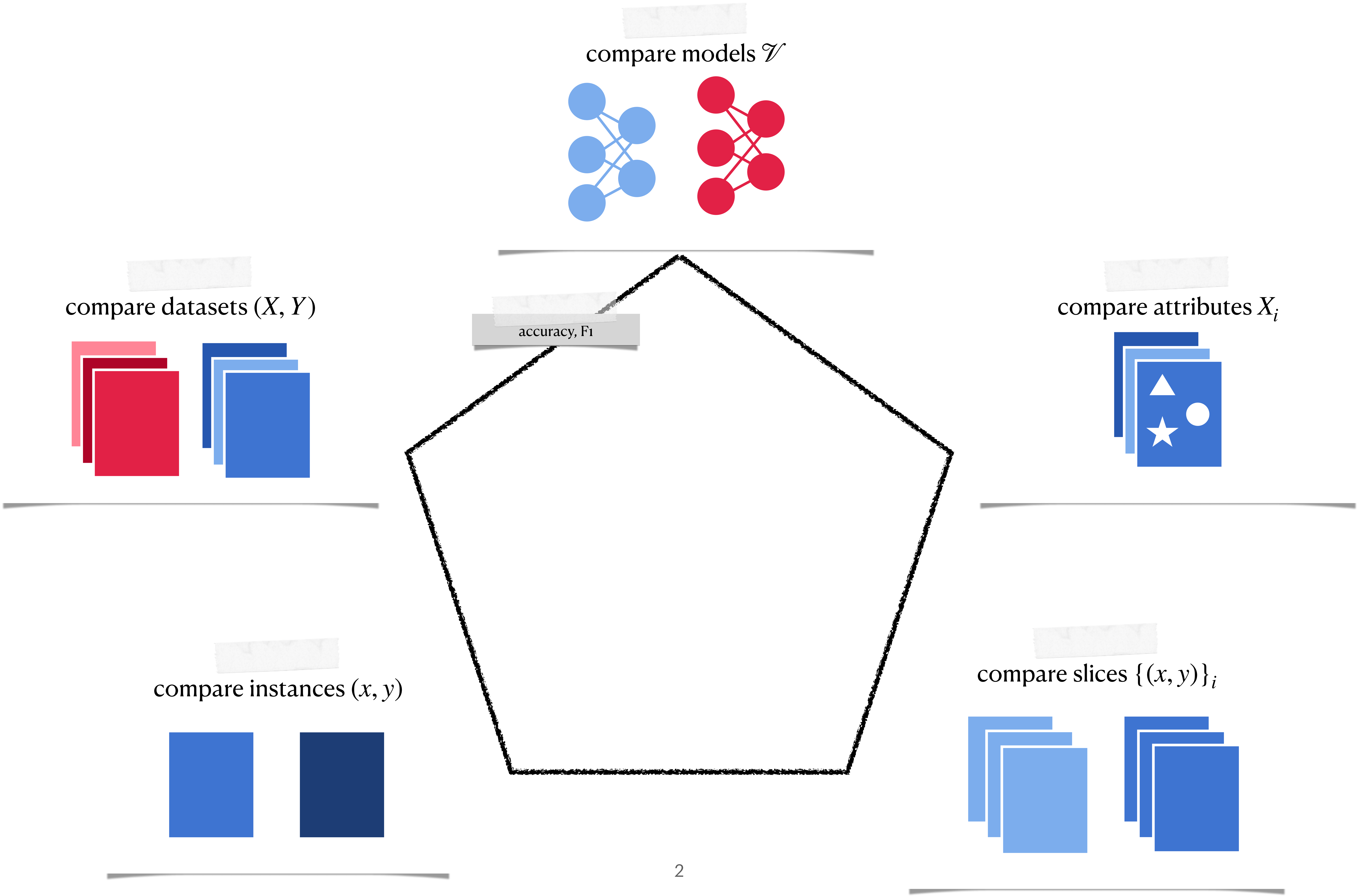


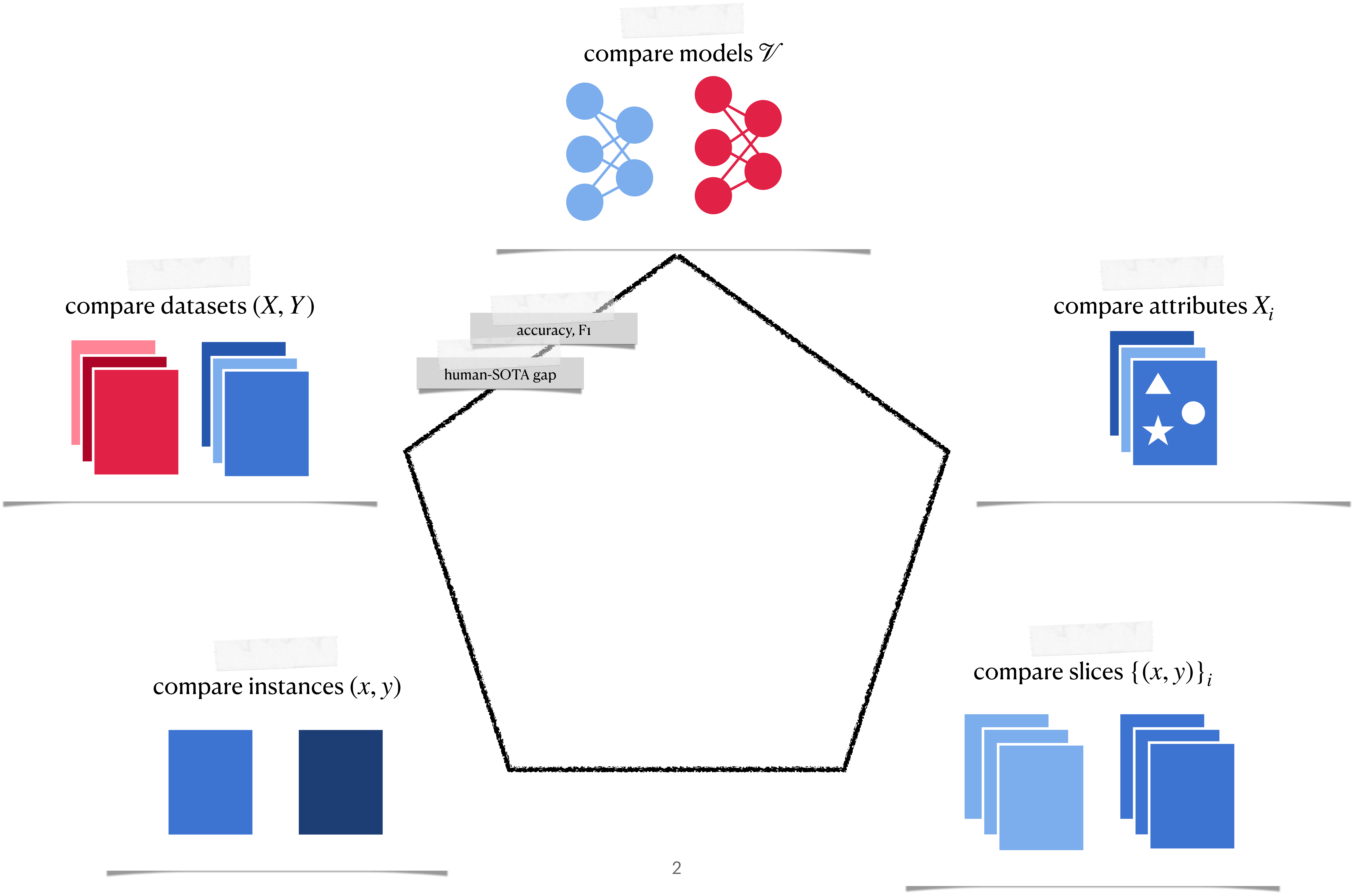
Kawin Ethayarajh — Stanford NLP

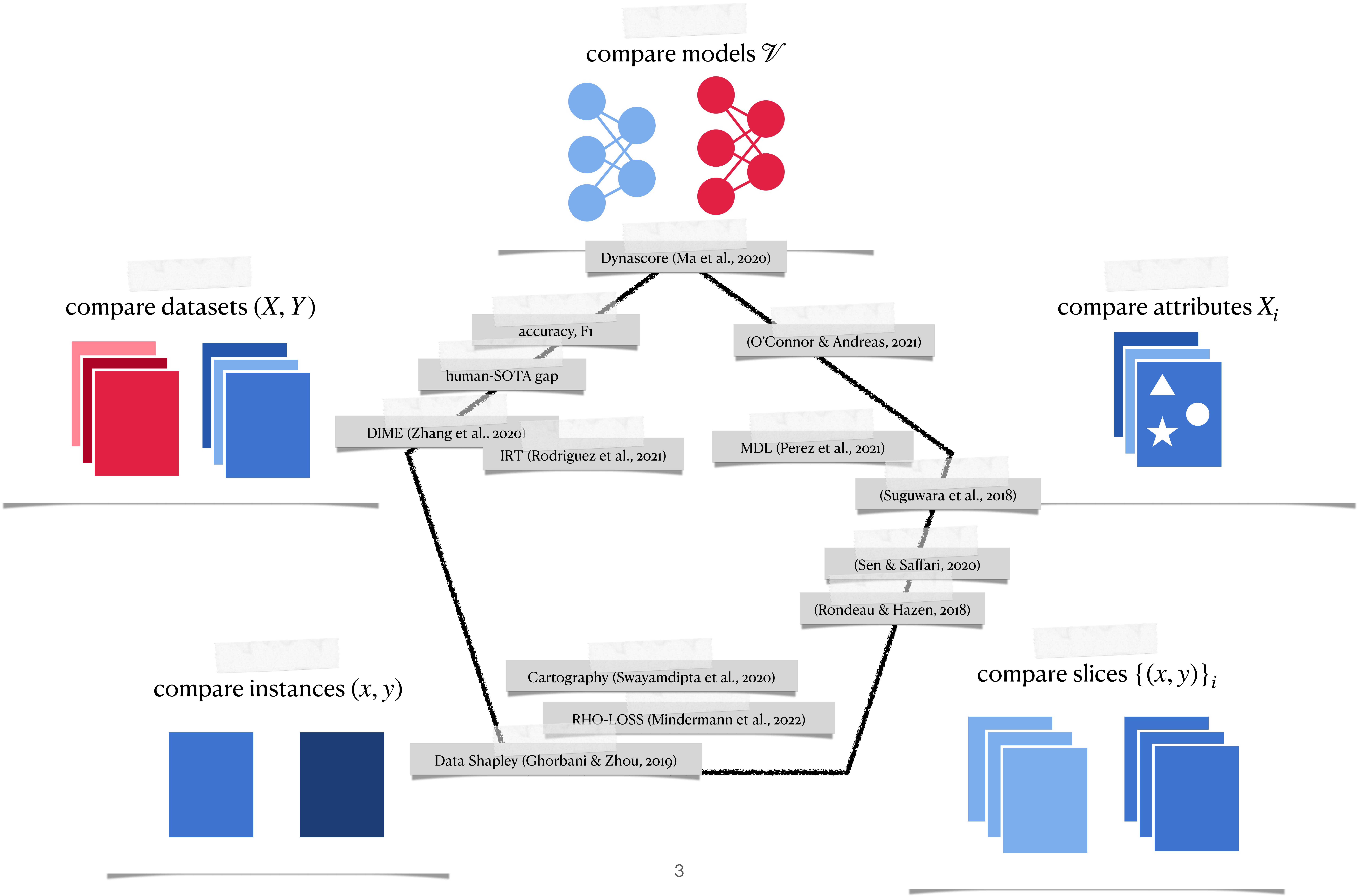
Yejin Choi — UWNLP, AI2

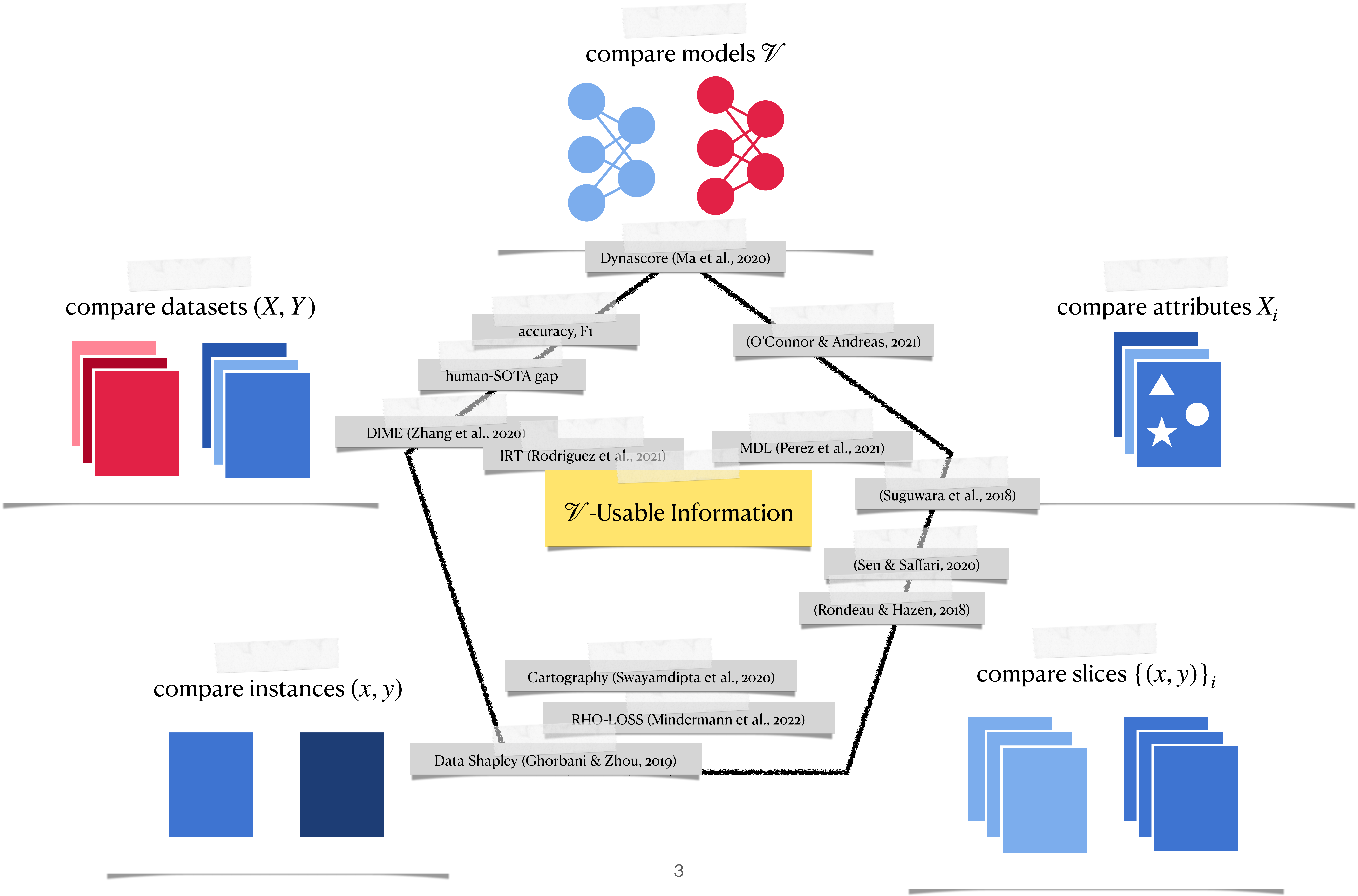
Swabha Swayamdipta — AI2


compare models $\mathscr{V}$
compare datasets (X, Y)
compare attributes X_i
compare instances (x, y)
compare slices $\{(x, y)\}_i$

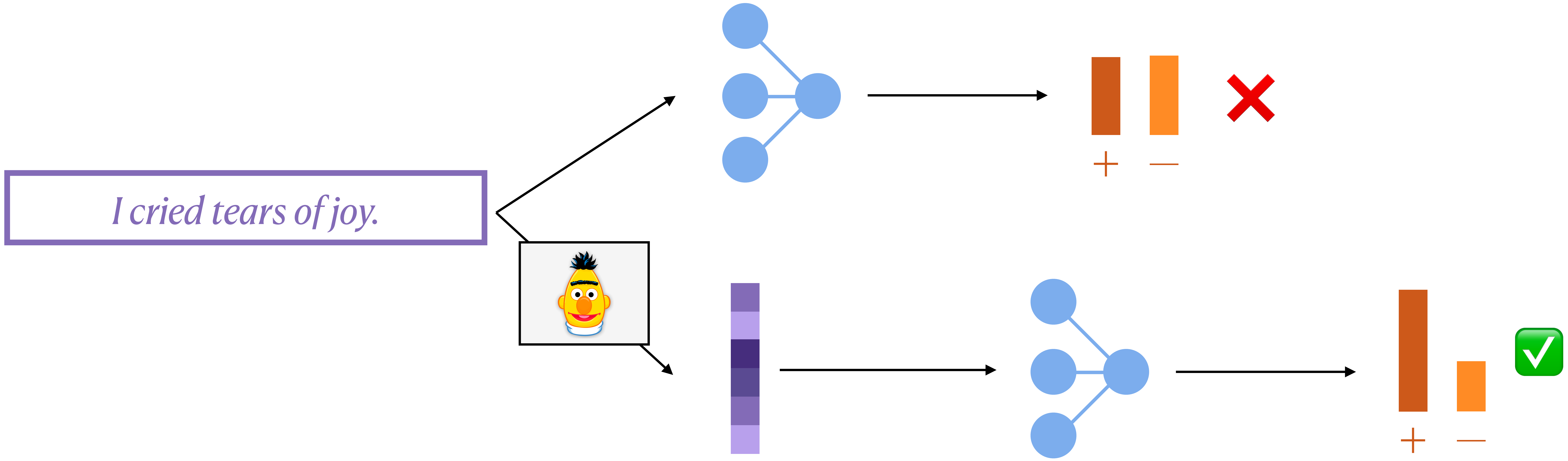

compare models $\mathscr{V}$
compare datasets (X, Y)
accuracy, F1
compare attributes X_i
compare instances (x, y)
compare slices $\{(x, y)\}_i$


compare models $\mathscr{V}$
compare datasets (X, Y)
compare attributes X_i
accuracy, F1
human-SOTA gap
compare instances (x, y)
compare slices $\{(x, y)\}_i$


compare models $\mathscr{V}$
Dynascore (Ma et al., 2020)
compare datasets (X, Y)
accuracy, F1
human-SOTA gap
DIME (Zhang et al., 2020)
IRT (Rodriguez et al., 2021)
(O'Connor & Andreas, 2021)
MDL (Perez et al., 2021)
compare attributes X_i
(Suguwara et al., 2018)
(Sen & Saffari, 2020)
(Rondeau & Hazen, 2018)
compare slices $\{(x, y)\}_i$
Cartography (Swayamdipta et al., 2020)
RHO-LOSS (Mindermann et al., 2022)
Data Shapley (Ghorbani & Zhou, 2019)
compare instances (x, y)


compare models $\mathcal{V}$
Dynascore (Ma et al., 2020)
accuracy, F1
human-SOTA gap
DIME (Zhang et al., 2020)
IRT (Rodriguez et al., 2021)
(O'Connor & Andreas, 2021)
MDL (Perez et al., 2021)
compare datasets (X, Y)
compare attributes X_i
$\mathcal{V}$-Usable Information
(Suguwara et al., 2018)
(Sen & Saffari, 2020)
(Rondeau & Hazen, 2018)
Cartography (Swayamdipta et al., 2020)
RHO-LOSS (Mindermann et al., 2022)
Data Shapley (Ghorbani & Zhou, 2019)
compare instances (x, y)
compare slices $\{(x, y)\}_i$

Transforming the input with τ can make information previously *unusable* by model family $\mathcal{V}$ now *usable*, despite $I(X; Y) \geq I(\tau(X); Y)$.



[Xu et al., 2019]

The predictive $\mathcal{V}$-information framework can be used to measure the amount of usable information X contains about Y w.r.t. $\mathcal{V}$.

$$I_{\mathcal{V}}(X \to Y) = \underbrace{\inf_{f \in \mathcal{V}} \mathbb{E}[-\log_2 f[\varnothing](Y)]}_{H_{\mathcal{V}}(Y)} - \underbrace{\inf_{f \in \mathcal{V}} \mathbb{E}[-\log_2 f[X](Y)]}_{H_{\mathcal{V}}(Y|X)}$$

[Xu et al., 2019]

The predictive $\mathcal{V}$-information framework can be used to measure the amount of usable information X contains about Y w.r.t. $\mathcal{V}$.

$$I_{\mathcal{V}}(X \to Y) = \underbrace{\inf_{f \in \mathcal{V}} \mathbb{E}[-\log_2 f[\varnothing](Y)]}_{H_{\mathcal{V}}(Y)} - \underbrace{\inf_{f \in \mathcal{V}} \mathbb{E}[-\log_2 f[X](Y)]}_{H_{\mathcal{V}}(Y|X)}$$

train/finetune on null input $\varnothing$

[Xu et al., 2019]

The predictive $\mathscr{V}$-information framework can be used to measure the amount of usable information X contains about Y w.r.t. $\mathscr{V}$.

$$I_{\mathscr{V}}(X \to Y) = \underbrace{\inf_{f \in \mathscr{V}} \mathbb{E}[-\log_2 f[\varnothing](Y)]}_{H_{\mathscr{V}}(Y)} - \underbrace{\inf_{f \in \mathscr{V}} \mathbb{E}[-\log_2 f[X](Y)]}_{H_{\mathscr{V}}(Y|X)}$$

train/finetune on null input $\varnothing$

train/finetune on actual input X

[Xu et al., 2019]

The predictive $\mathscr{V}$-information framework can be used to measure the amount of usable information X contains about Y w.r.t. $\mathscr{V}$.

$I_{\mathscr{V}}(X$ - ... $X](Y)]$

The lower the $\mathscr{V}$-usable information, the more difficult the dataset is for $\mathscr{V}$.

train/finetune on null input $\varnothing$ train/finetune on actual input X

[Xu et al., 2019]

SNLI

[Bowman et al., 2015]

natural language inference

PREMISE: Women enjoying a game of table tennis.

HYPOTHESIS: Women enjoying a game of ping pong.

- ● entailment
- ○ neutral
- ○ contradiction

MultiNLI

[Williams et al., 2018]

natural language inference

PREMISE: The Old One always comforted Ca'daan, except today.

HYPOTHESIS: Ca'daan knew the Old One very well.

- ○ entailment
- ● neutral
- ○ contradiction

CoLA

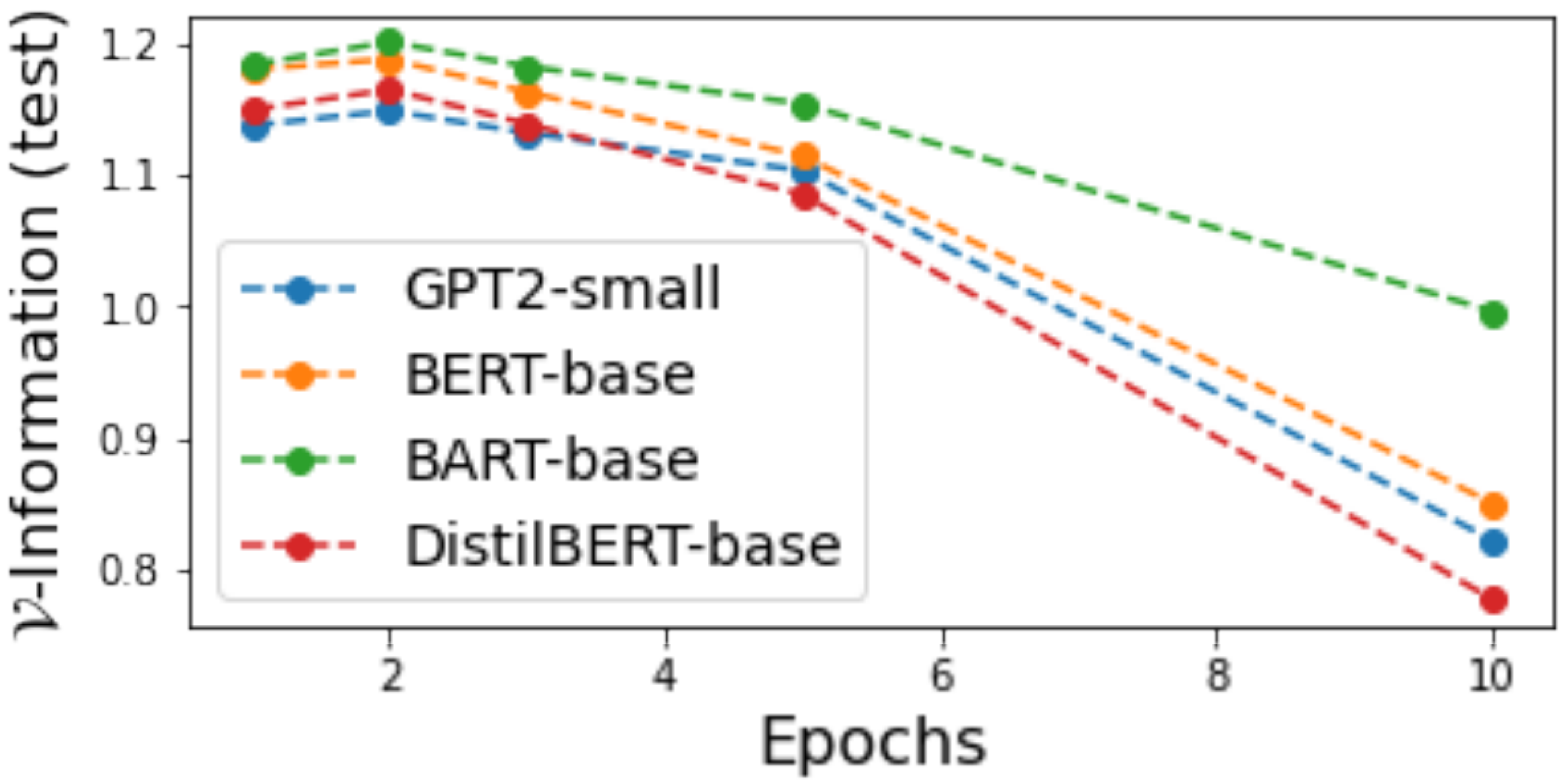
[Warstadt et al., 2018]

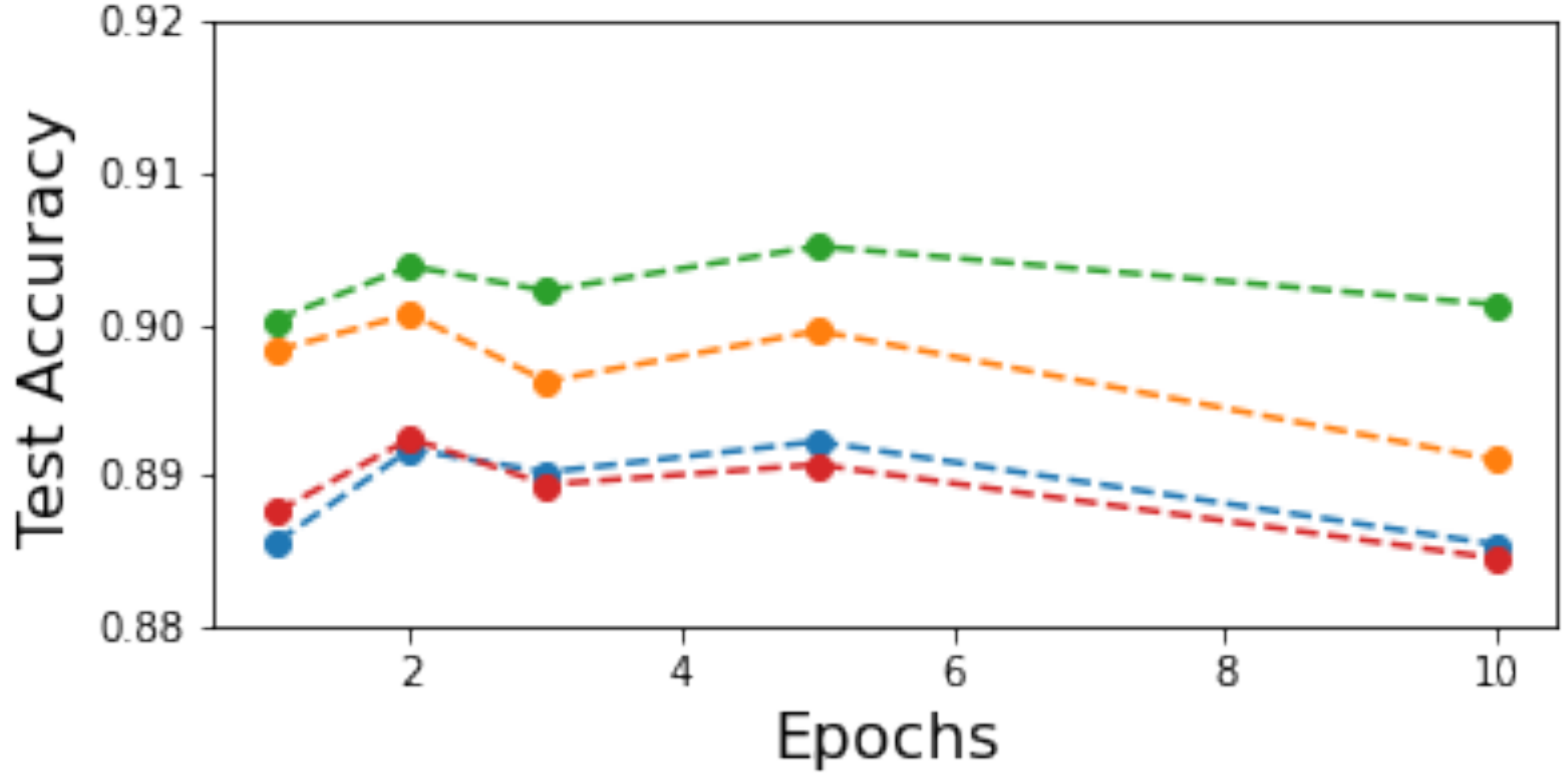
text classification

Wash you.

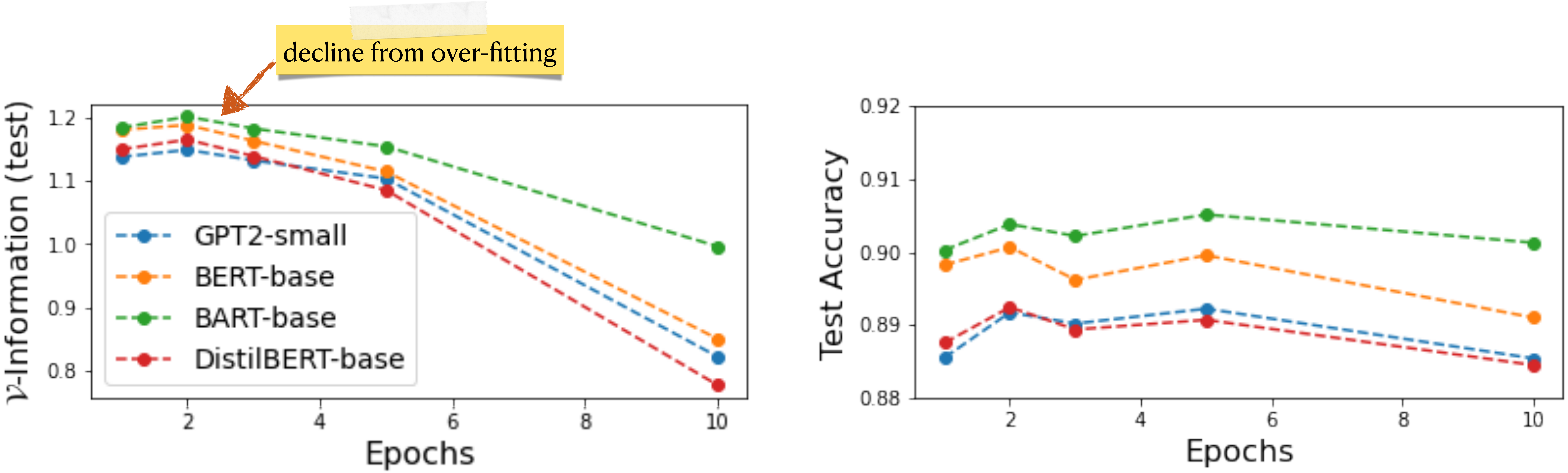
- ○ grammatical
- ● ungrammatical

Compare different models $\mathscr{V}$ by computing $I_{\mathscr{V}}(X \to Y)$ for the same (X, Y), shown here for SNLI.

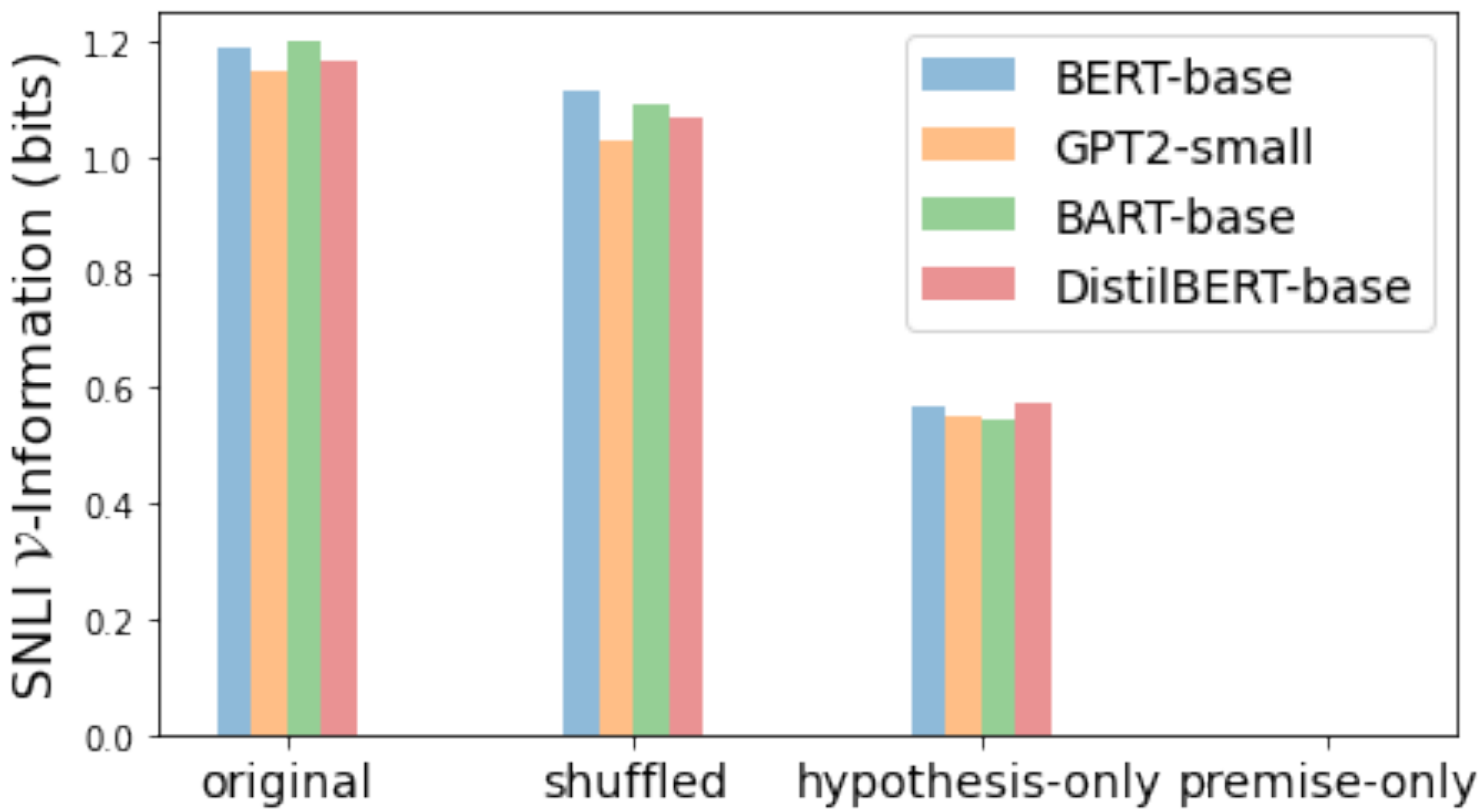




Compare different models $\mathcal{V}$ by computing $I_{\mathcal{V}}(X \to Y)$ for the same (X, Y), shown here for SNLI.

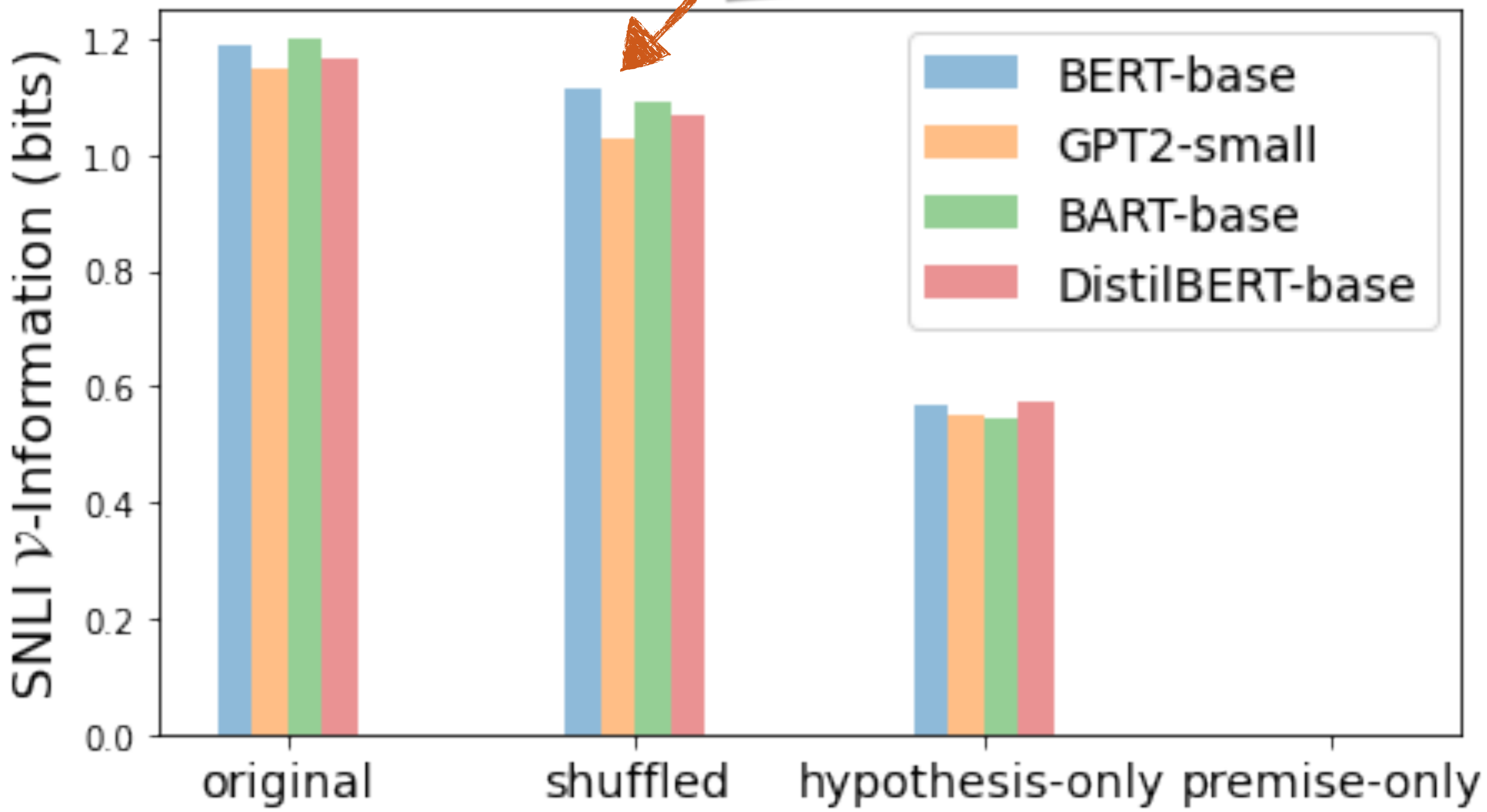


Compare different input attributes X_i by computing $I_{\mathcal{V}}(X_i \to Y)$ for the same $Y, \mathcal{V}$.

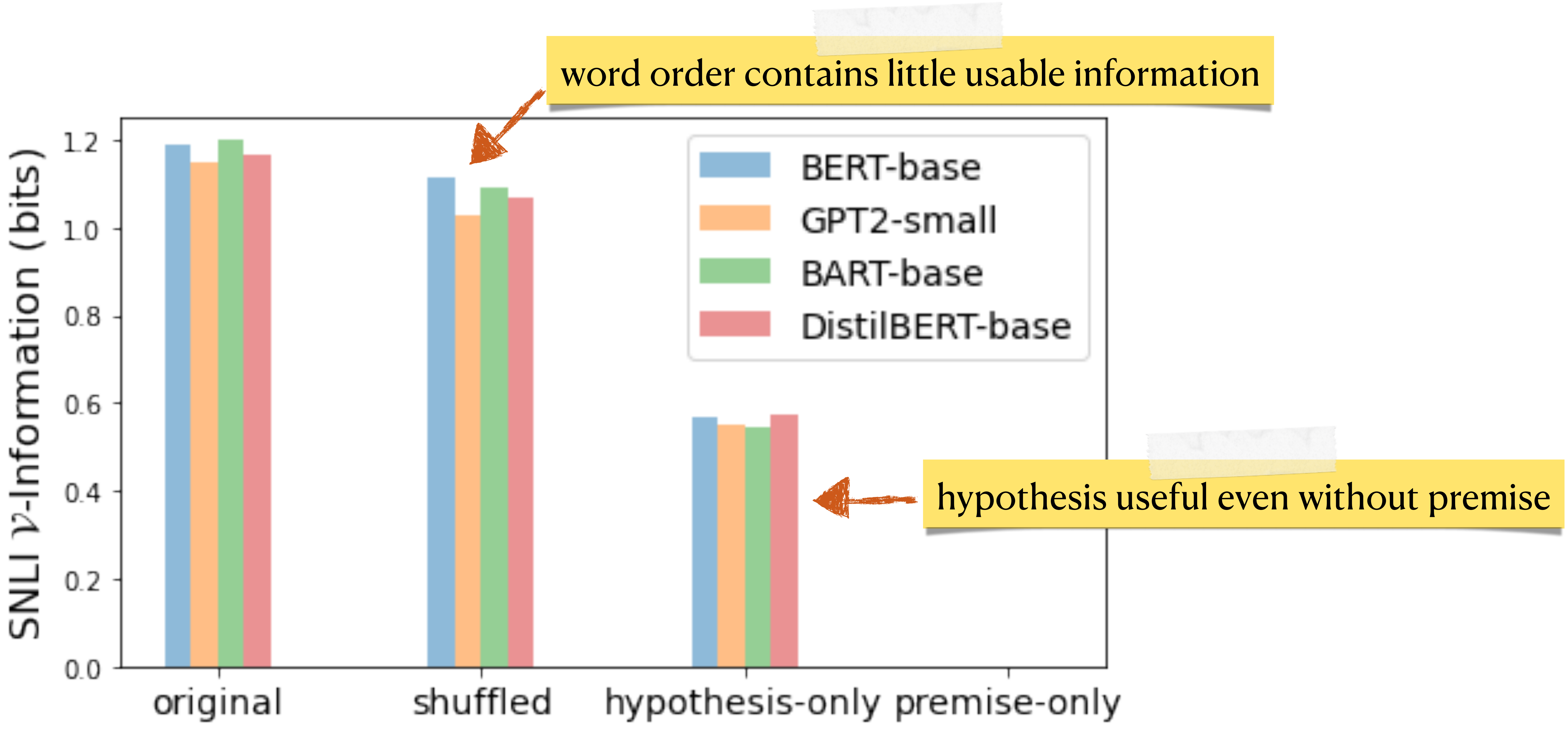


Compare **different input attributes X_i** by computing $I_{\mathcal{V}}(X_i \to Y)$ for the same $Y, \mathcal{V}$.

word order contains little usable information



Compare different input attributes X_i by computing $I_{\mathcal{V}}(X_i \to Y)$ for the same $Y, \mathcal{V}$.



We can measure instance-level difficulty (w.r.t. a distribution) with pointwise $\mathcal{V}$-information (PVI), the analogue of PMI.

$$I_{\mathcal{V}}(X \to Y) = \mathbb{E}_{x,y \sim P(X,Y)}[\text{PVI}(x \to y)]$$

We can measure instance-level difficulty (w.r.t. a distribution) with pointwise $\mathcal{V}$-information (PVI), the analogue of PMI.

$$I_{\mathcal{V}}(X \to Y) = \mathbb{E}_{x,y \sim P(X,Y)}[\text{PVI}(x \to y)]$$

$I_{\mathcal{V}}(X \to Y) \in \mathbb{R}^{0+}$; $\text{PVI}(x \to y) \in \mathbb{R}$

We can measure instance-level difficulty (w.r.t. a distribution) with pointwise $\mathcal{V}$-information (PVI), the analogue of PMI.

$$I_{\mathcal{V}}(X \to Y) = \mathbb{E}_{x,y \sim P(X,Y)}[\text{PVI}(x \to y)]$$

$I_{\mathcal{V}}(X \to Y) \in \mathbb{R}^{0+}$; $\text{PVI}(x \to y) \in \mathbb{R}$

cross-epoch Pearson's $r \geq 0.747$

We can measure instance-level difficulty (w.r.t. a distribution) with pointwise $\mathscr{V}$-information (PVI), the analogue of PMI.

$$I_{\mathscr{V}}(X \to Y) = \mathbb{E}_{x,y \sim P(X,Y)}[\mathrm{PVI}(x \to y)]$$

$I_{\mathscr{V}}(X \to Y) \in \mathbb{R}^{0+}$; $\mathrm{PVI}(x \to y) \in \mathbb{R}$

cross-epoch Pearson's $r \geq 0.747$

cross-seed Pearson's $r \geq 0.877$

We can measure instance-level difficulty (w.r.t. a distribution) with pointwise $\mathscr{V}$-inf

The higher the PVI, the easier the instance is for $\mathscr{V}$ w.r.t. $P(X, Y)$.

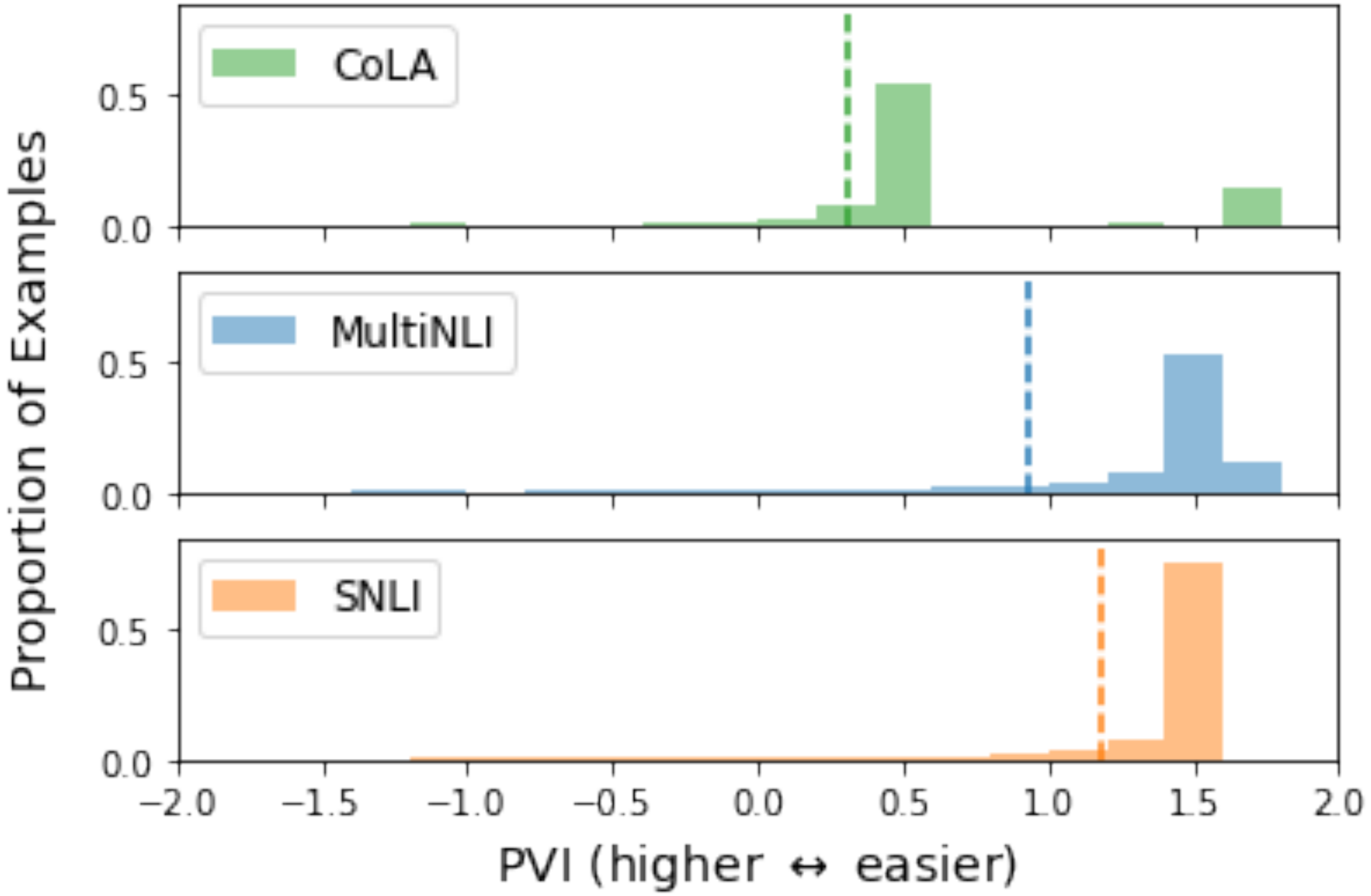
$$I_{\mathscr{V}}(X \to Y) = \mathbb{E}_{x,y \sim P(X,Y)}[\mathrm{PVI}(x \to y)]$$

$I_{\mathscr{V}}(X \to Y) \in \mathbb{R}^{0+}$; $\mathrm{PVI}(x \to y) \in \mathbb{R}$

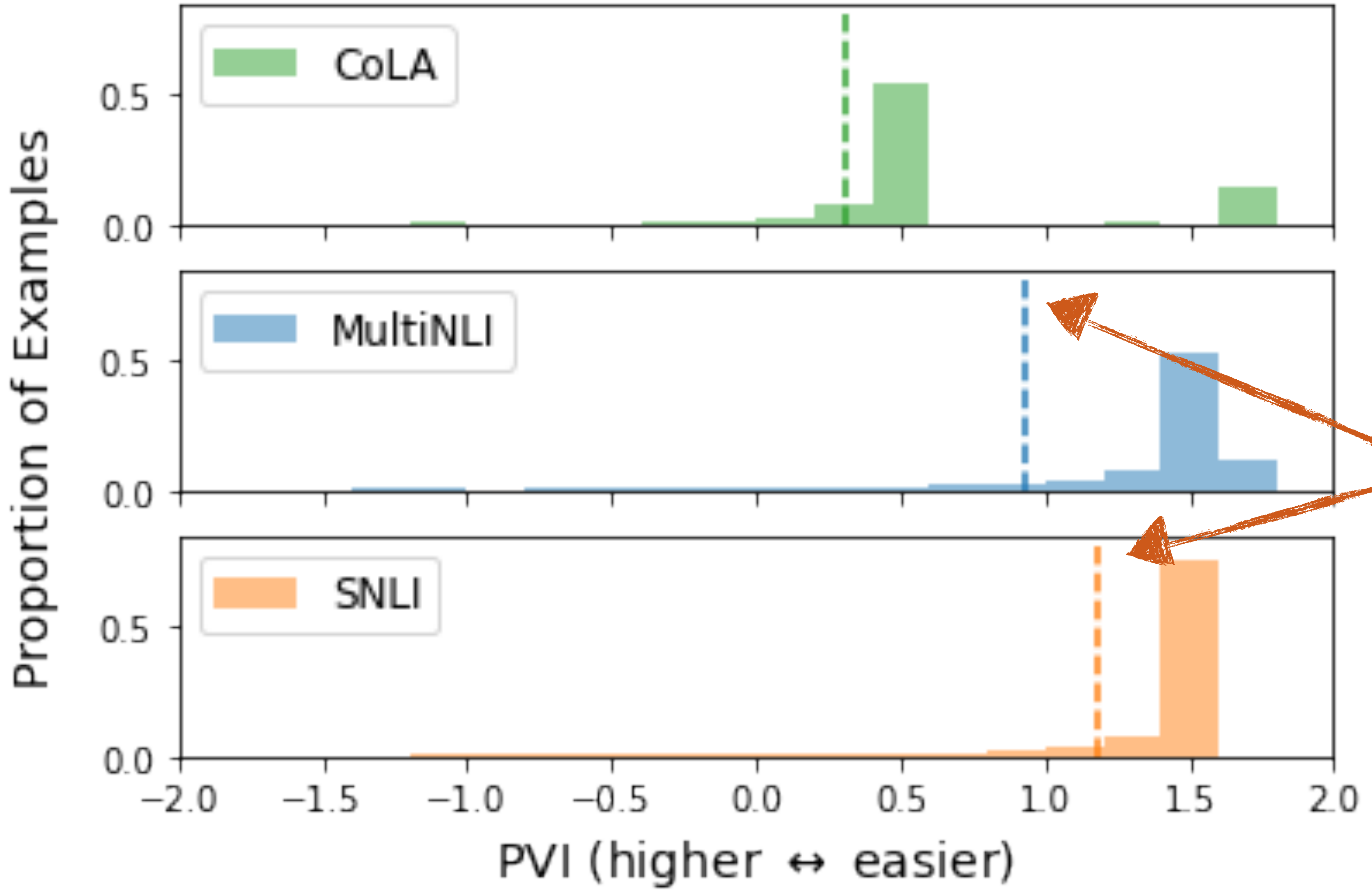
cross-epoch Pearson's $r \geq 0.747$

cross-seed Pearson's $r \geq 0.877$

Compare different datasets (X, Y) by estimating $I_{\mathcal{V}}(X \to Y)$ and PVI$(x \to y)$ for the same $\mathcal{V}$ across datasets.



Compare **different datasets** (X, Y) by estimating $I_{\mathcal{V}}(X \to Y)$ and PVI$(x \to y)$ for the same $\mathcal{V}$ across datasets.



same task, different dataset, different difficulty

Compare different instances (x, y) using PVI$(x \rightarrow y)$ for the same $\mathcal{V}$, X, Y, before and after transformations.

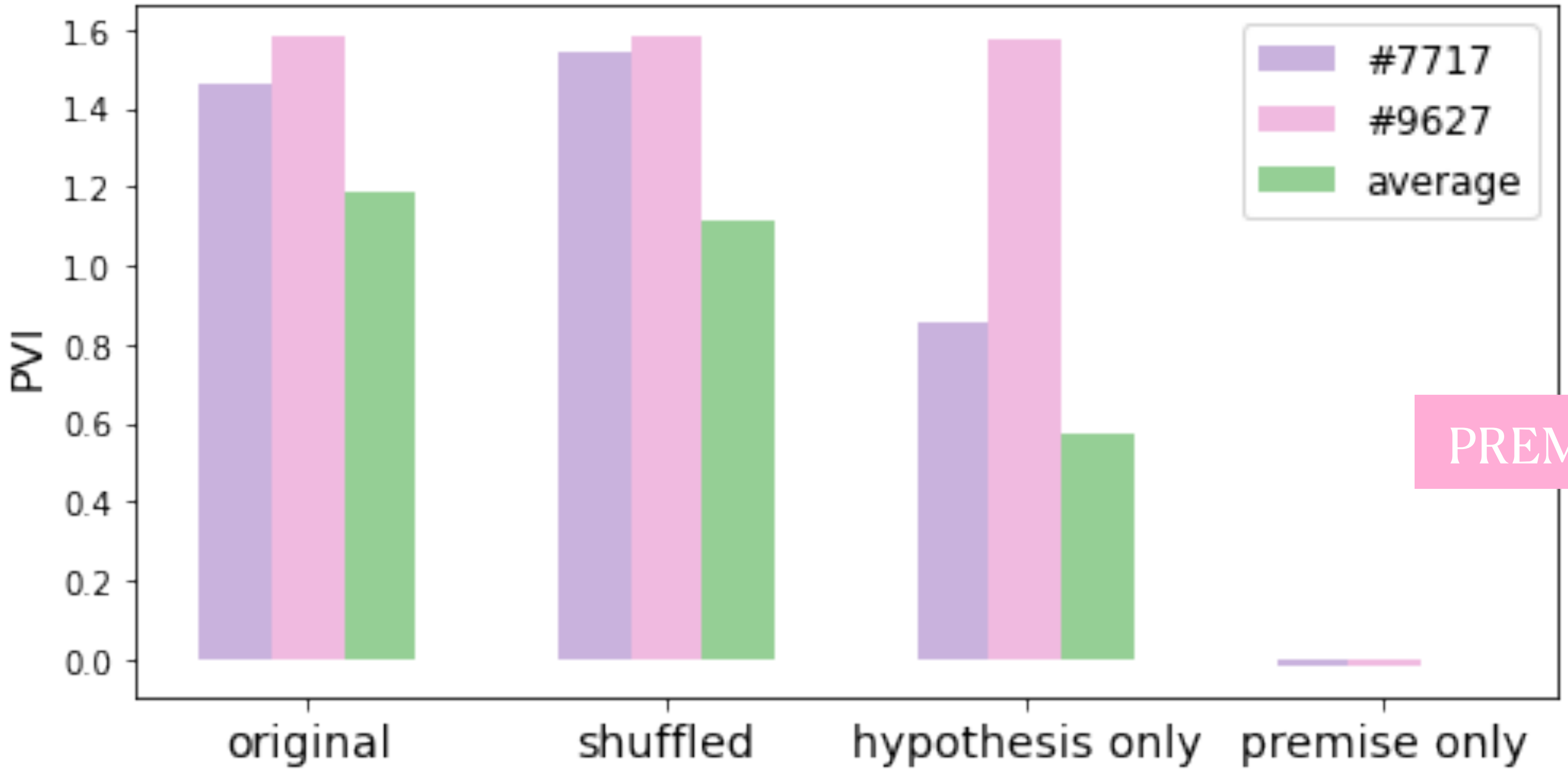
PREMISE: Little kids play a game of running around a pole.

HYPOTHESIS: The kids are fighting outside.

PREMISE: A group of people watching a boy getting interviewed by a man.

HYPOTHESIS: A group of people are sleeping on Pluto.

Compare different instances (x, y) using $\text{PVI}(x \to y)$ for the same $\mathcal{V}$, X, Y, before and after transformations.



PREMISE: Little kids play a game of running around a pole.

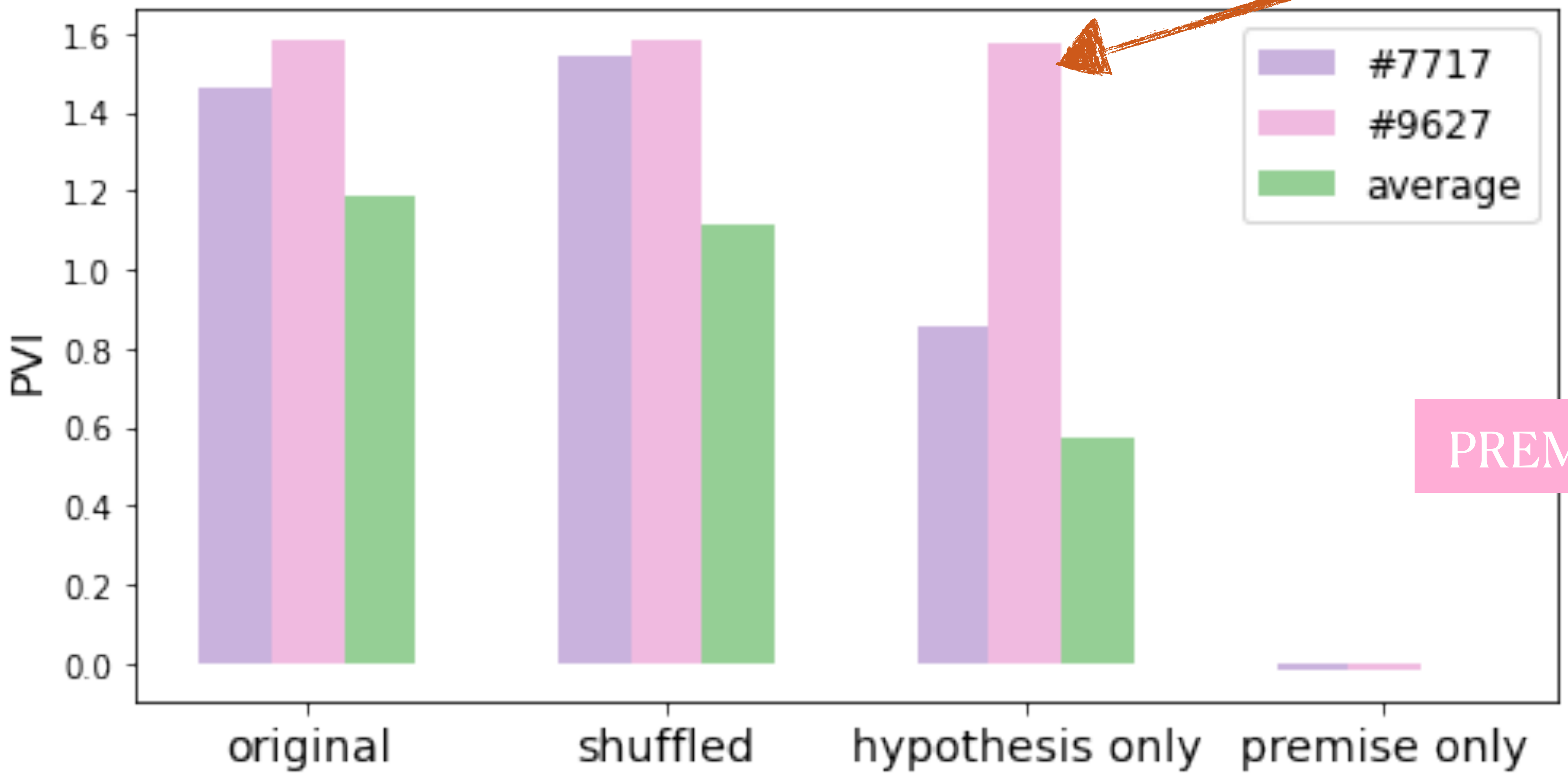
HYPOTHESIS: The kids are fighting outside.

PREMISE: A group of people watching a boy getting interviewed by a man.

HYPOTHESIS: A group of people are sleeping on Pluto.

Compare different instances (x, y) using PVI$(x \rightarrow y)$ for the same $\mathcal{V}$, X, Y, before and after transformations.

hypothesis is what makes #9627 easier!



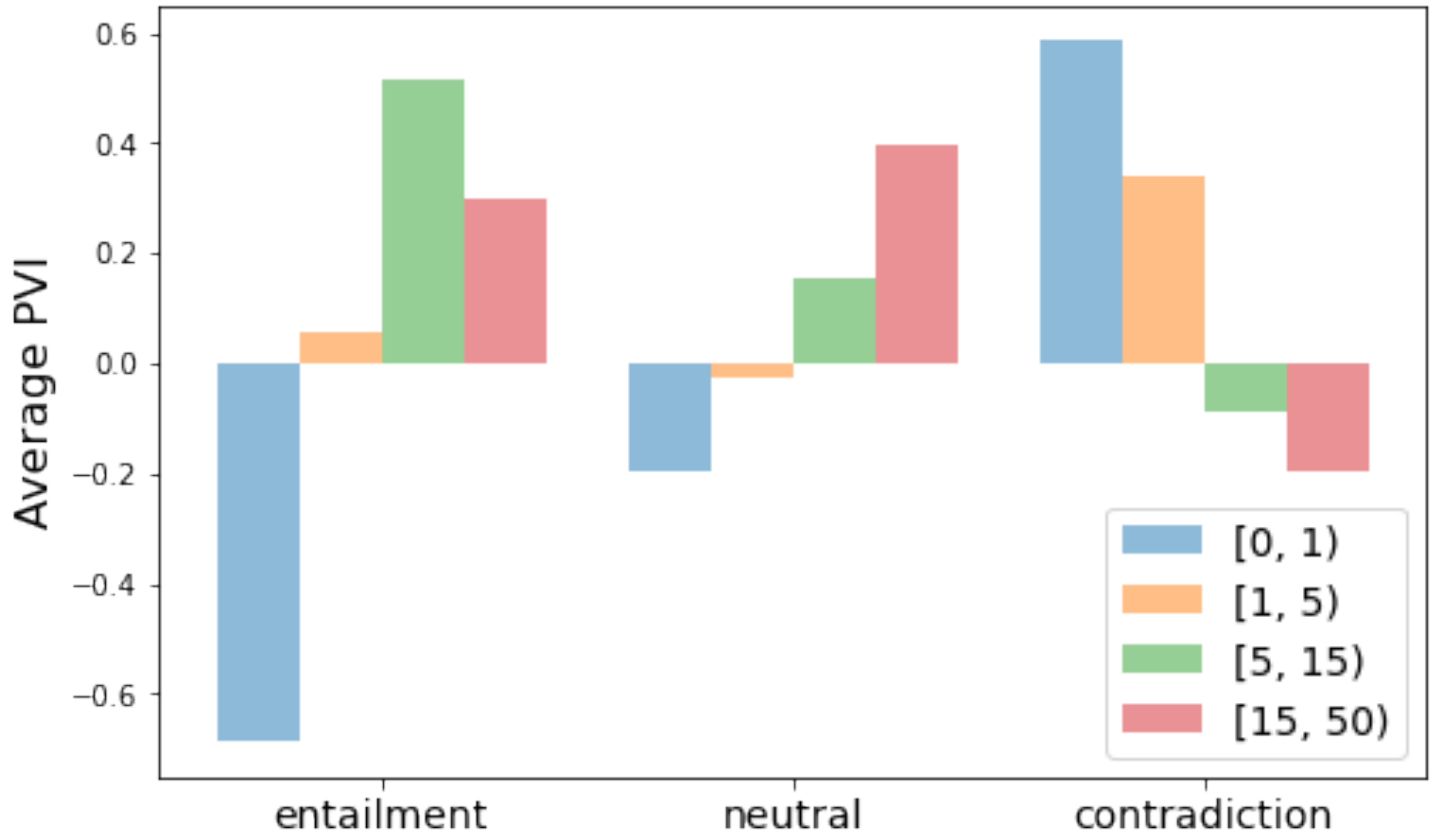
PREMISE: Little kids play a game of running around a pole.

HYPOTHESIS: The kids are fighting outside.

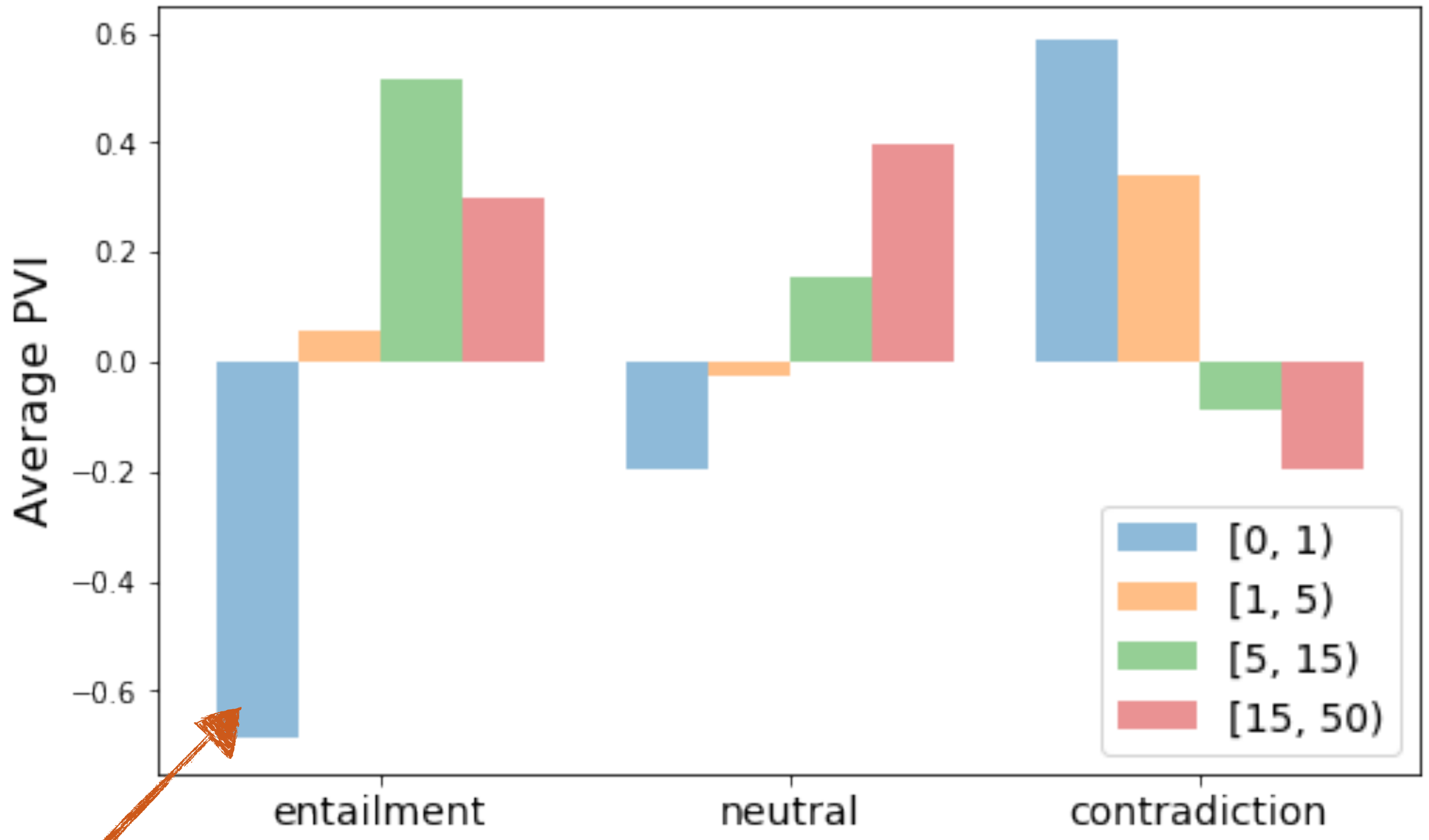
PREMISE: A group of people watching a boy getting interviewed by a man.

HYPOTHESIS: A group of people are sleeping on Pluto.

Compare different slices $\{(x, y)\}_i$ by estimating the average PVI$(x \rightarrow y)$ for each slice.



Compare different slices $\{(x, y)\}_i$ by estimating the average PVI$(x \rightarrow y)$ for each slice.



what BERT finds hardest!

Estimating the drop in $\mathcal{V}$-information after leaving out a token reveals token-level annotation artefacts.

Grammatical (CoLA)

will	0.267
John	0.168
.	0.006
and	-0.039
in	-0.050

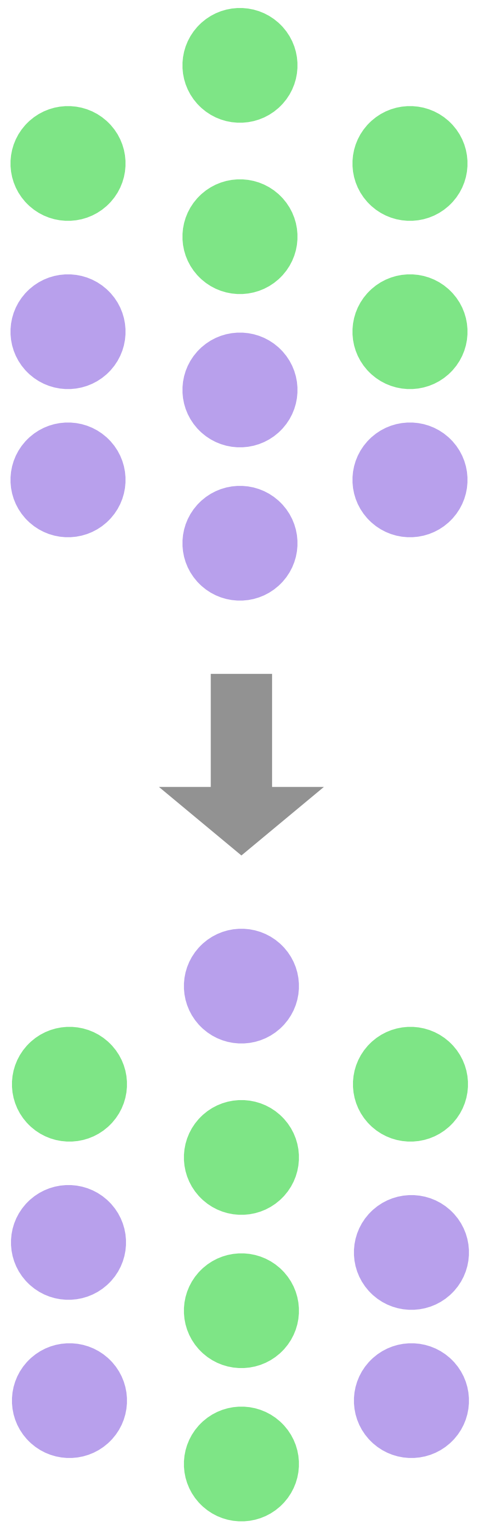
Ungrammatical (CoLA)

book	2.737
is	2.659
was	2.312
of	2.308
in	1.972

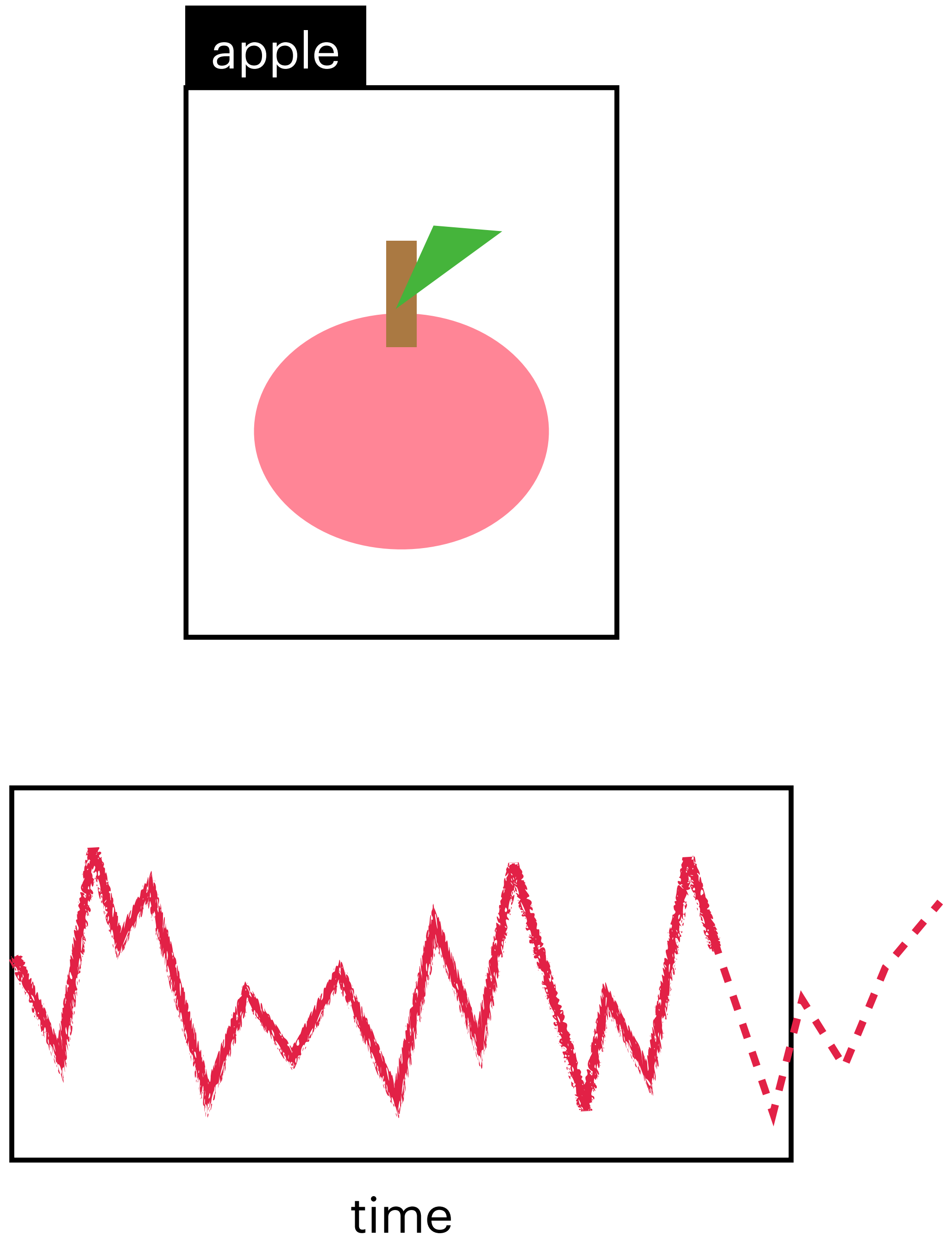
[Gururangan et al., 2018]

Future Work

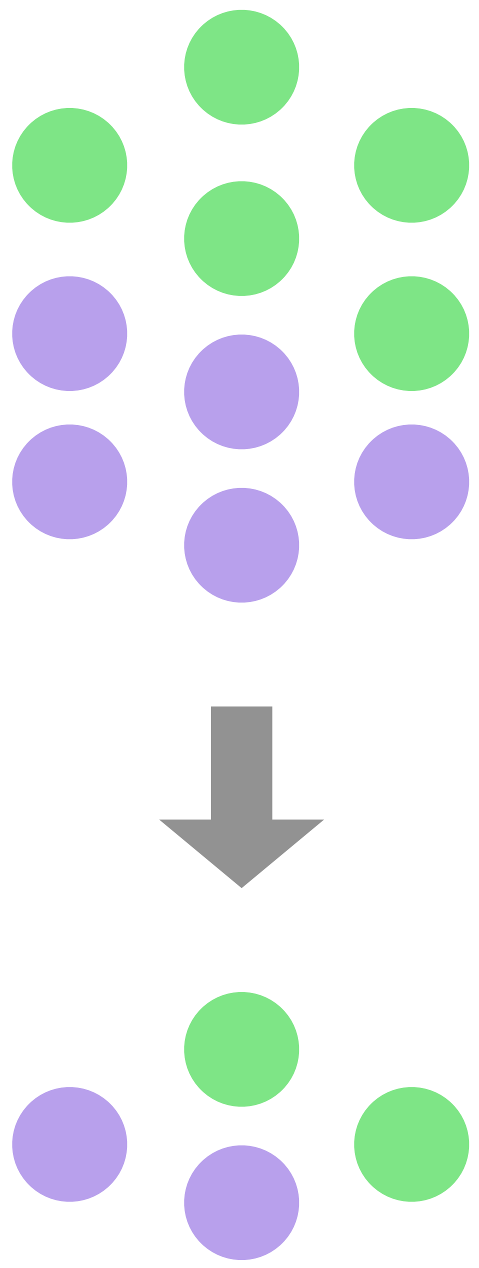
Making Tougher Datasets



Other Modalities



Data Pruning



Summary: A unified framework for interpreting datasets.